Прикладная статистика
2. Тесты о равенстве средних.

Родионов Игорь Владимирович
rodionov@bigdatateam.org

Параметры сдвига и масштаба

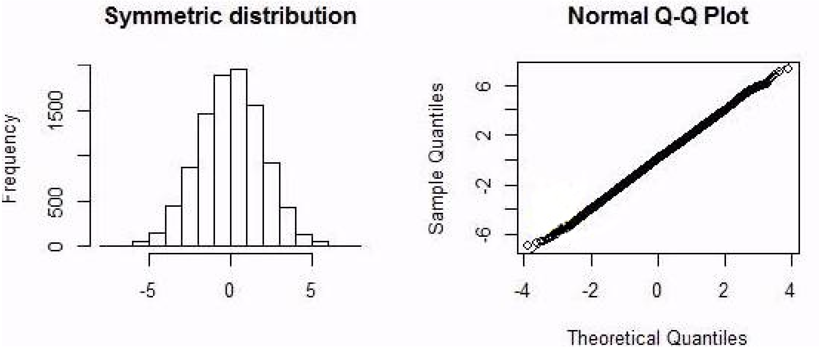
Пусть имеется параметрическое семейство функций распределения $\{F_\theta(x), \theta \in \Theta\}$.

- Параметр θ называется параметром сдвига, если $F_\theta(x) = F_0(x - \theta)$;
- Параметр θ называется параметром масштаба, если $F_\theta(x) = F_1(x/\theta)$;
- Параметр, не являющийся параметром сдвига или масштаба, называется параметром формы.

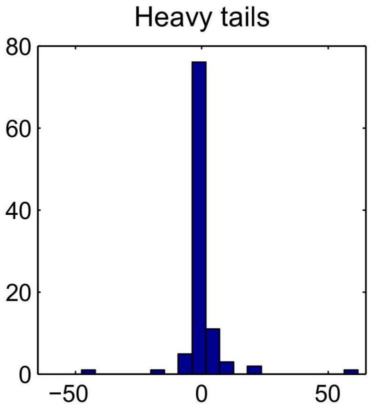
К примеру, для семейства распределений $N(a, \sigma^2)$ параметр a является параметром сдвига, а σ – параметром масштаба, оба параметра бета-распределения являются параметрами формы.

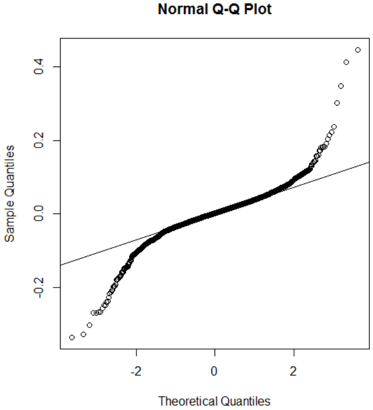
Допустим, мы хотим проверить гипотезу $H_0 : F = F_0\left(\frac{x-a}{\sigma}\right)$ (т.е. гипотеза о принадлежности семейству распределений с параметрами сдвига и масштаба).

QQ-plot – это график, на который нанесены точки $\left(F_0^{-1}\left(\frac{i-0.5}{n}\right), X_{(i)}\right)$. Если точки примерно лежат на одной прямой, то распределение данных близко к F_0 (с точностью до параметров сдвига и масштаба).

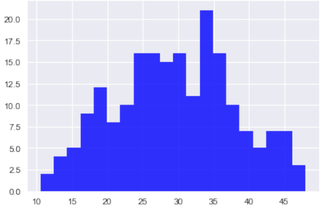


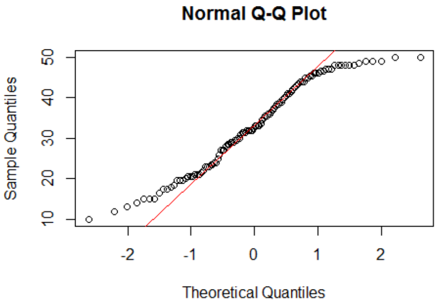
На рисунке: QQ-plot относительно нормального распределения (Normal QQ-plot) для выборки из нормального закона (среднее – 0, дисперсия $\approx$ 4). Так же QQ-plot может выглядеть для выборки из распределения Стьюдента с большим числом степеней свободы.



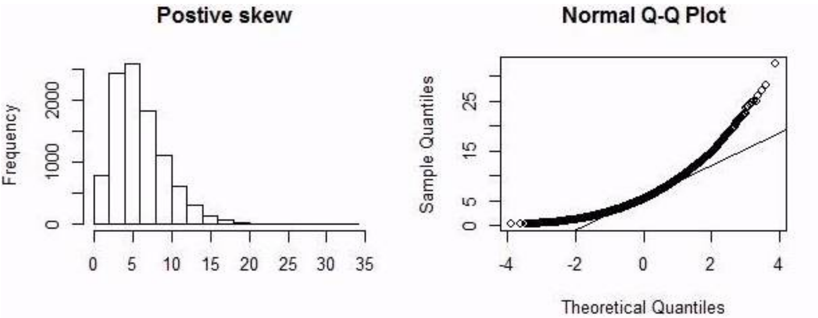


На рисунке: Normal QQ-plot для выборки с тяжелыми (относительно нормального закона) хвостами (heavy tails). Примеры: *Exp*, *Pareto*, *Gamma*, *LN*, *Cauchy* и тд.

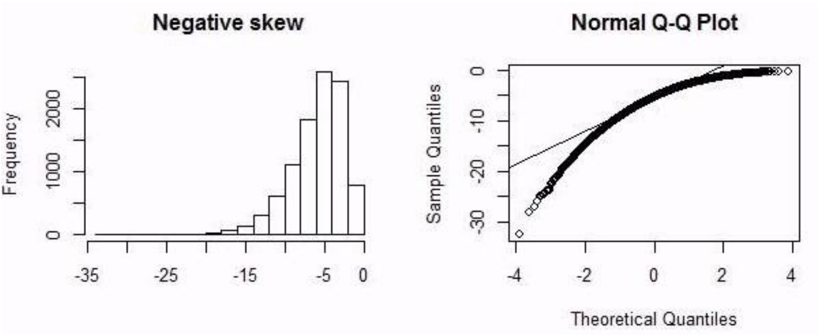




На рисунке: Normal QQ-plot для выборки с легкими (относительно нормального закона) хвостами (light tails). Примеры: *Weibull*(α), $\alpha > 2$, *Gumbel*, распределения с ограниченным носителем.



На рисунке: Normal QQ-plot для выборки из правостороннего распределения (right-skewed). Примеры: *Gamma*, *Poisson*, *Beta*(α, β), $\alpha < \beta$.



На рисунке: Normal QQ-plot для выборки из левостороннего распределения (left-skewed). Примеры: *Inverse-Gamma*, *Beta*(α, β), $\alpha > \beta$.

Пусть $X_1, \ldots, X_n$ – выборка из неизвестного распределения с функцией распределения F. Рассмотрим гипотезу $H_0 : F = F_0$ против альтернативы $H_1 : F \neq F_0$. Критерии проверки таких гипотез называются **критериями согласия**.

Примеры:

1) Критерий Колмогорова-Смирнова (KS-test)

если $\sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F_0(x)| > K_{1-\alpha}$, *то отвергаем* H_0,

здесь $K_{1-\alpha}$ – квантиль распределения Колмогорова.

2) Критерий Андерсона-Дарлинга (Ω^2-критерий) со статистикой

$$\Omega^2 = \int \frac{(\widehat{F}_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x).$$

3) Критерий хи-квадрат Пирсона.
Пусть P_0 – вероятностная мера (на $\mathbb{R}$), соответствующая функции распределения F_0. Разобьем $\mathbb{R}$ на $6-10$ интервалов $\{B_i\}_{i=1}^k$ таких, что $nP_0(B_i) \geq 5\ \forall i$ (B_i выбираются заранее, вне зависимости от данных!).
Обозначим $p_i^0 = P_0(B_i)$, $\mu_i = \sharp\{j : X_j \in B_i\}$. Обозначим

$$\widehat{\chi} := \sum_{i=1}^{k} \frac{(\mu_i - np_i^0)^2}{np_i^0}.$$

Критерий хи-квадрат выглядит так

$$\textit{если } \widehat{\chi} > u_{1-\alpha}, \textit{ то отвергнуть } H_0,$$

где $u_{1-\alpha}$ – $(1-\alpha)$-квантиль распределения χ^2_{k-1}.

Модификация критериев согласия

Один из методов проверки принадлежности распределения выборки какому-либо семейству распределений заключается в применении критериев согласия.

Пусть семейство функций распределения $\{F(x; a, b)\}$ параметризовано параметрами сдвига и масштаба, т.е. $F(x; a, b) = F_0(\frac{x-a}{b})$, где $F_0(x) = F(x; 0, 1)$.

Пусть $\widehat{a}$ и $\widehat{b}$ – состоятельные оценки параметров a и b соответственно.

Тогда для проверки гипотезы $H_0 : F \in \{F(x; a, b)\}$ можно воспользоваться критерием согласия, где вместо F_0 в статистике критерия нужно подставить $F_0(\frac{x-\widehat{a}}{\widehat{b}})$.

Критерий Лиллиефорса

В частности, если для семейства нормальных распределений $\{N(a, \sigma^2)\}$ в качестве критерия согласия взять критерий Колмогорова-Смирнова со статистикой $D_n = \sup_x |\widehat{F}_n(x) - F_0(x)|$, а в качестве оценок параметров a и σ^2 – выборочное среднее и выборочную дисперсию соответственно, то получим **критерий Лиллиефорса** проверки нормальности выборки.

Замечание. Предельное распределение статистик модифицированных критериев согласия при верности нулевой гипотезы могут существенно зависеть не только от вида семейства распределений, но и от вида выбранных оценок параметров.

Проверка нормальности

Критерий Шапиро-Уилка
Проверяется гипотеза $H_0 : F \in \{N(a, \sigma^2)\}$ против альтернативы $H_1 : F \notin \{N(a, \sigma^2)\}$. Статистика критерия

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i-1}^{n} (X_i - \overline{X})^2},$$

где $(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{\|m^T V^{-1}\|^2}$, $(m_1, \ldots, m_n)$ – мат. ожидания порядковых статистик выборки размера n из $N(0, 1)$, а V – матрица ковариаций этих порядковых статистик.

При верной гипотезе H_0 статистика W имеет табличное распределение, значения вектора $(a_1, \ldots, a_n)$ также табулированы.

Проверка нормальности

Критерий Харке-Бера тоже используется для проверки выборки на нормальность. Рассмотрим статистику

$$JB = \frac{n}{6}((Sk)^2 + \frac{1}{4}(Ku)^2),$$

где $Sk = \frac{m_3}{m_2^{3/2}}$, $Ku = \frac{m_4}{m_2^2} - 3$ и $m_j = \frac{1}{n}\sum_i (X_i - \overline{X})^j$ – выборочный центральный момент.

При $n > 2000$ $JB \approx \chi^2(2)$, при меньших n следует искать квантили в таблицах. Модификацией критерия Харке-Бера является довольно точный комбинированный K^2-критерий.

Сравнение критериев проверки нормальности распределения случайных величин

Наименование критерия (раздел)	Характер альтернативного распределения					Ранг
	асимметричное		симметричное		≈ нормальное	
	$\alpha_4 < 3$	$\alpha_4 > 3$	$\alpha_4 < 3$	$\alpha_4 > 3$	$\alpha_4 \approx 3$	
Критерий Шапиро–Уилка (3.2.2.1)	1	1	3	2	2	1
Критерий K^2 (3.2.2.16)	7	8	10	6	4	2
Критерий Дарбина (3.1.2.7)	11	7	7	15	1	3
Критерий Д'Агостино (3.2.2.14)	12	9	4	5	12	4
Критерий α_4 (3.2.2.16)	14	5	2	4	18	5
Критерий Васичека (3.2.2.2)	2	14	8	10	10	6
Критерий Дэвида–Хартли–Пирсона (3.2.2.10)	21	2	1	9	1	7
Критерий χ^2 (3.1.1.1)	9	20	9	8	3	8
Критерий Андерсона–Дарлинга (3.1.2.4)	18	3	5	18	7	9
Критерий Филлибена (3.2.2.5)	3	12	18	1	9	10
Критерий Колмогорова–Смирнова (3.1.2.1)	16	10	6	16	5	11
Критерий Мартинеса–Иглевича (3.2.2.14)	10	16	13	3	15	12
Критерий Лина–Мудхолкара (3.2.2.13)	4	15	12	12	16	13
Критерий α_3 (3.2.2.16)	8	6	21	7	19	14
Критерий Шпигельхальтера (3.2.2.11)	19	13	11	11	8	15
Критерий Саркади (3.2.2.12)	5	18	15	14	13	16
Критерий Смирнова–Крамера–фон Мизеса (3.1.2.2)	17	11	20	17	6	17
Критерий Локка–Спурье (3.2.2.7)	13	4	19	21	17	18
Критерий Оя (3.2.2.8)	20	17	14	13	14	19
Критерий Хегази–Грина (3.2.2.3)	6	19	16	19	21	20
Критерий Муроты–Такеучи (3.2.2.17)	15	21	17	20	20	21

Кобзарь, 3.2.2.19, табл. 80.

t-критерий Стьюдента

Предположим, что $(X_1, \ldots, X_n)$ – выборка из $N(\mu, \sigma^2)$. Проверим гипотезу $H_0 : \mu = \mu_0$ против альтернативы $H_1 : \mu \neq \mu_0$. Если

$$\left| \sqrt{n} \frac{\overline{X} - \mu_0}{s} \right| > t_{1-\alpha/2},$$

то отвергаем H_0 на уровне значимости α. Здесь $t_{1-\alpha/2}$ – квантиль распределения Стьюдента с $n-1$ степенью свободы, s – корень из выборочной дисперсии.

Если неожиданно оказалось, что мы знаем дисперсию выборки, то в критерии s заменяется на σ, квантиль – на квантиль $N(0, 1)$ того же уровня, а сам тест будет называться Z-тестом.

Двухвыборочный t-тест

Предположим, что $(X_1, \ldots, X_n) \sim N(\mu_1, \sigma^2)$, $(Y_1, \ldots, Y_m) \sim N(\mu_2, \sigma^2)$, т.е. дисперсии распределений одинаковы, причем σ неизвестна, и что выборки независимы. Проверим гипотезу $H_0 : \mu_1 = \mu_2$ против альтернативы $H_1 : \mu_1 \neq \mu_2$. Если

$$\left| \sqrt{\frac{nm}{n+m}} \cdot \frac{\overline{X} - \overline{Y}}{S} \right| > t_{1-\alpha/2},$$

то отвергаем H_0 на уровне значимости α. Здесь $t_{1-\alpha/2}$ – квантиль распределения Стьюдента с $n+m-2$ степенью свободы, а

$$S = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}}.$$

Но как проверить, что дисперсии двух нормальных выборок равны? Для этого воспользуемся критерием Фишера.

Пусть $(X_1, \ldots, X_n) \sim N(\mu_1, \sigma_1^2)$, $(Y_1, \ldots, Y_m) \sim N(\mu_2, \sigma_2^2)$, выборки независимы. Проверим гипотезу $H_0 : \sigma_1 = \sigma_2$ против $H_1 : \sigma_1 \neq \sigma_2$. Если

$$\frac{s_X^2}{s_Y^2} \notin (u_{\alpha/2}, u_{1-\alpha/2}),$$

то отвергаем H_0. Здесь u_γ – квантиль распределения Фишера с $n-1$ и $m-1$ степенями свободы.

Критерий Фишера, как и t-критерии, чувствителен к отклонению выборок от нормального распределения.

Что делать, если выяснилось, что дисперсии выборок различны? Пусть, как и ранее, $(X_1, \ldots, X_n) \sim N(\mu_1, \sigma_1^2)$, $(Y_1, \ldots, Y_m) \sim N(\mu_2, \sigma_2^2)$, выборки независимы. Тогда при верной гипотезе $H_0 : \mu_1 = \mu_2$ статистика

$$\frac{\overline{X} - \overline{Y}}{\sqrt{s_X^2/n + s_Y^2/m}}$$

приблизительно распределена по закону Стьюдента с K степенями свободы, где

$$K = \left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2 \cdot \left(\frac{s_X^4}{n^2(n-1)} + \frac{s_Y^4}{m^2(m-1)}\right)^{-1} - 2.$$

А что делать, если выборки оказались зависимыми? В этом случае проверку гипотезы $H_0 : \mu_1 = \mu_2$ можно свести к одномерной.

Пусть $(X_1, \ldots, X_n) \sim N(\mu_1, \sigma^2)$, $(Y_1, \ldots, Y_n) \sim N(\mu_2, \sigma^2)$, выборки зависимы и одинакового размера. При верной гипотезе H_0 выполнено

$$\sqrt{n} \cdot \frac{\overline{X} - \overline{Y}}{S} \sim St(n-1),$$

где $D_i = X_i - Y_i$ и

$$S = \frac{1}{\sqrt{n-1}} \sqrt{\sum_{i=1}^{n} (D_i - \overline{D})^2}.$$

Одновыборочный критерий знаков

Обсудим теперь ситуацию, когда выборки не являются нормальными. Проще всего в таком случае перейти к непараметрическим тестам.

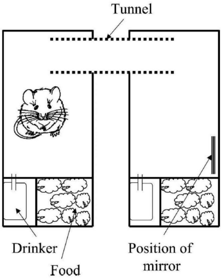
Имеем выборку $X = (X_1, \ldots, X_n)$, где $X_i \neq m_0$. Хотим проверить гипотезу $H_0 : med(X) = m_0$ против $H_1 : med(X) \neq m_0$, где $med(X)$ – медиана распределения выборки X.

Используем для этого статистику $T(X) = \sum_i I(X_i > m_0)$. Ясно, что при верной гипотезе H_0

$$T \sim Bin(n, 1/2).$$

Одновыборочный критерий знаков

Пример: (Shervin, 2004) 16 лабораторных мышей были помещены в двухкомнатные клетки, в одной из комнат висело зеркало. Измерялось доля времени, которое каждая мышь проводила в каждой из своих двух клеток.

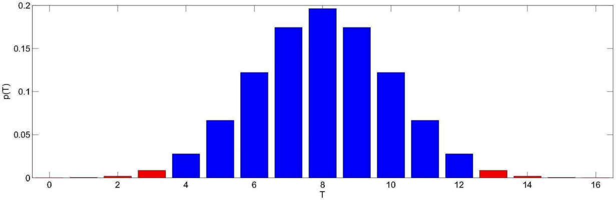


Постановка задачи:
H_0 : мышам всё равно, висит в клетке зеркало или нет.
H_1 : у мышей есть какие-то предпочтения насчёт зеркала.

Одновыборочный критерий знаков

Будем действовать в рамках такой постановки: H_0 : медиана доли времени, проводимого в клетке с зеркалом, равна $1/2$, H_1 : медиана доли времени, проводимого в клетке с зеркалом, не равна $1/2$.

Редуцированные данные: 0 – мышь провела больше времени в комнате с зеркалом, 1 – в комнате без зеркала.
Статистика: T – число единиц в выборке.



13 из 16 мышей провели больше времени в комнате без зеркала.

Критерий знаков: $p = 0.0213$, 95% доверительный интервал для вероятности (что мышь проведёт больше времени в комнате без зеркала) – $[0.54, 0.96]$.

Двухвыборочный критерий знаков

Пусть зависимость между выборками $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_n)$ неизвестна, причем (желательно) $X_i \neq Y_i \ \forall i$. Потребуем, чтобы распределение выборки величин $\{X_i - Y_i\}_{i=1}^n$ было симметричным.

Хотим проверить гипотезу об отсутствии эффекта
$H_0 : P(X > Y) = 1/2$ против альтернативы
$H_1 : P(X > Y) \neq 1/2$.

Для проверки гипотезы используем статистику
$T(X, Y) = \sum_i I(X_i > Y_i)$. Ясно, что при верной H_0

$$T(X, Y) \sim Bin(n, 1/2).$$

Критерий резонно использовать, когда 1) точные разности $X_i - Y_i$ неизвестны, известны только их знаки; 2) разности небольшие по модулю, но систематические по знаку; 3) разности большие по модулю, но случайные по знаку.

Одновыборочный критерий знаковых рангов Уилкоксона

Более точным по сравнению с критерием знаков является критерий Уилкоксона, однако он требует, чтобы распределение выборки было симметричным.

Имеем выборку $X = (X_1, \ldots, X_n)$ из симметричного распределения, где $X_i \neq m_0$. Хотим проверить гипотезу $H_0 : med(X) = m_0$ против $H_1 : med(X) \neq m_0$.

Рассмотрим статистику

$$W = \sum_i rank(|X_i - m_0|) sign(X_i - m_0).$$

Тогда при верной гипотезе H_0 статистика W имеет табличное распределение. При $n > 20$ (и верной H_0) можно использовать аппроксимацию

$$W \sim N(0, n(n+1)(2n+1)/6).$$

Двухвыборочный критерий знаковых рангов Уилкоксона

Пусть зависимость между выборками $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_n)$ неизвестна, причем (желательно) $X_i \neq Y_i\ \forall i$. Потребуем, чтобы распределение выборки величин $\{V_i\}_{i=1}^n$, где $V_i = X_i - Y_i$ было симметричным.

Для проверки гипотезы $H_0 : med(X - Y) = 0$ рассмотрим статистику

$$T = \sum_i rank(|V_i|)I(V_i > 0).$$

Тогда при верной гипотезе H_0 статистика W имеет табличное распределение. При $n > 20$ (и верной H_0) можно использовать аппроксимацию

$$T \sim N(n(n+1)/4, n(n+1)(2n+1)/24).$$

Замечания о знаковых тестах

1) При $n < 20$ пользоваться нормальной аппроксимацией не рекомендуется, следует использовать табличные значения квантилей.

2) Чтобы проверить симметричность распределения, можно нанести на график точки (ξ_i, η_i), где $\xi_i = -V_{(i)} + \widehat{\mu}$, $\eta_i = V_{(n-i+1)} - \widehat{\mu}$, $i = 1, \ldots, [n/2]$, $\widehat{\mu}$ – выборочная медиана. Точки должны приближаться прямой $y = x$.

3) Если некоторые V_i (или другие разности) равны 0, то отбрасываем их и уменьшаем n.

4) Если среди $|V_i|$ есть совпадения, то следует использовать средние ранги (дисперсия нормального приближения тоже изменится).

5) Вместо статистики T в критерии знаковых рангов Уилкоксона может использоваться статистика $W = \sum_i rank(|V_i|) sign(V_i)$.

Критерий Манна-Уитни-Уилкоксона

Пусть $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_m)$ – 2 независимые выборки с функциями распределения F_X и F_Y соответственно, причем $F_X(x - \theta) = F_Y(x)$ и $n < m$. Проверим гипотезу об отсутствии сдвига $H_0 : \theta = 0$ против $H_1 : \theta \neq 0$.

Составим вариационный ряд объединенной совокупности $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ и обозначим через R_i ранг наблюдения X_i в этом вариационном ряду. Определим статистику Манна-Уитни

$$W = \sum_{i=1}^{n} R_i.$$

Тогда при верной H_0 W имеет табличное распределение.

Критерий Манна-Уитни-Уилкоксона

1) При $n, m > 20$ и верности гипотезы H_0 можно пользоваться нормальной аппроксимацией

$$W \sim N\left(\frac{n(n+m+1)}{2}, \frac{nm(n+m+1)}{2}\right).$$

В остальных случаях стоит пользоваться табличными значениями квантилей критерия.

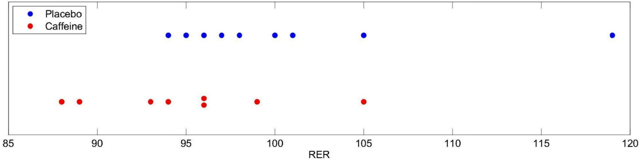
2) Если некоторые наблюдения совпадают (т.е. им присвоены средние ранги), то в аппроксимации следует заменить дисперсию на следующую:

$$\frac{mn(m+n+1)}{2}\left(1 - \frac{\sum_{i=1}^{k} t_i(t_i^2 - 1)}{(m+n)((m+n)^2 - 1)}\right),$$

где k – число групп совпавших величин и t_i – число величин в i-той группе.

Критерий Манна-Уитни-Уилкоксона

Пример. RER (респираторный обмен) – соотношение числа молекул CO_2 и O_2 в выдыхаемом воздухе. Является косвенным признаком того, из жиров или углеводов вырабатывается энергия в момент измерения. Изучалось влияние кофеина на мышечный метаболизм. В эксперименте принимало участие 18 испытуемых, респираторный обмен которых измерялся в процессе физических упражнений. За час до этого 9 из них получили таблетку кофеина, оставшиеся 9 – плацебо. Повлиял ли кофеин на значение показателя респираторного обмена?



H_0 : среднее значение RER не отличается в двух группах.
H_1 : среднее значение RER отличается в двух группах.

Критерий Манна-Уитни-Уилкоксона

Ранг	Наблюдение	Номер наблюдения	Наблюдение	Ранг
16.5	105	1	96	9
18	119	2	99	13
14	100	3	94	5.5
11	97	4	89	3
9	96	5	96	9
15	101	6	93	4
5.5	94	7	88	1.5
7	95	8	105	16.5
12	98	9	88	1.5

Статистика W – сумма рангов одной из групп (т.к. $n = m = 9$, то неважно, какой).

$p = 0.0521$, 95% доверительный интервал для медианной разности – $[-0.00005, 1.2]$.

Спасибо за внимание!